Catena 151 (2017) 147–160

Contents lists available at ScienceDirect

Catena

journal homepage: www.elsevier.com/locate/catena

A comparative study of logistic model tree, random forest, and classification and regression tree models for spatial prediction of landslide susceptibility

Wei Chen [a], Xiaoshen Xie [a], Jiale Wang [a], Biswajeet Pradhan [b,c], Haoyuan Hong [d,*], Dieu Tien Bui [e], Zhao Duan [a], Jianquan Ma [a]

[a] College of Geology & Environment, Xi'an University of Science and Technology, Xi'an 710054, China
[b] Department of Civil Engineering, Faculty of Engineering, University Putra Malaysia, 43400, Selangor, Malaysia
[c] Department of Energy and Mineral Resources Engineering, Choongmu-gwan, Sejong University, 209 Neungdong-ro Gwangjin-gu, Seoul, 05006, Republic of Korea
[d] Jiangxi Provincial Meteorological Observatory, Jiangxi Meteorological Bureau, No.109 ShengfuBeier Road, Nanchang 330046, China
[e] Geographic Information System group, Department of Business Administration and Computer Science, University College of Southeast, Bø i Telemark N-3800, Norway

ARTICLE INFO

Article history:
Received 28 May 2016
Received in revised form 24 November 2016
Accepted 29 November 2016
Available online xxxx

Keywords:
Logistic model tree
Random forest
Classification and regression tree
Landslide
China

ABSTRACT

The main purpose of the present study is to use three state-of-the-art data mining techniques, namely, logistic model tree (LMT), random forest (RF), and classification and regression tree (CART) models, to map landslide susceptibility. Long County was selected as the study area. First, a landslide inventory map was constructed using history reports, interpretation of aerial photographs, and extensive field surveys. A total of 171 landslide locations were identified in the study area. Twelve landslide-related parameters were considered for landslide susceptibility mapping, including slope angle, slope aspect, plan curvature, profile curvature, altitude, NDVI, land use, distance to faults, distance to roads, distance to rivers, lithology, and rainfall. The 171 landslides were randomly separated into two groups with a 70/30 ratio for training and validation purposes, and different ratios of non-landslides to landslides grid cells were used to obtain the highest classification accuracy. The linear support vector machine algorithm (LSVM) was used to evaluate the predictive capability of the 12 landslide conditioning factors. Second, LMT, RF, and CART models were constructed using training data. Finally, the applied models were validated and compared using receiver operating characteristics (ROC), and predictive accuracy (ACC) methods. Overall, all three models exhibit reasonably good performances; the RF model exhibits the highest predictive capability compared with the LMT and CART models. The RF model, with a success rate of 0.837 and a prediction rate of 0.781, is a promising technique for landslide susceptibility mapping. Therefore, these three models are useful tools for spatial prediction of landslide susceptibility.

1. Introduction

Landslides, as one of the most commonly geological hazards in the world, cause thousands of casualties and fatalities, hundreds of billions of dollars in damage, and environmental losses each year (Aleotti and Chowdhury, 1999; Gutiérrez et al., 2015). For China, many regions have been seriously affected by landslide occurrences, and landslides have caused serious threats to the environment, settlements, and industrial facilities in the recent years (Lin et al., 2012; Ma et al., 2015; Wang et al., 2015a; Xu et al., 2015; Xu et al., 2014; Zhou et al., 2013).

Generally, landslide damages can be decreased to a certain extent by predicting future landslide locations (Pradhan, 2010). Globally, several statistical models combined with GIS have been used for landslide susceptibility assessment, such as statistical index (Chen et al., 2016a; Constantin et al., 2011; Nasiri Aghdam et al., 2016), index of entropy (IOE) (Constantin et al., 2011; Devkota et al., 2013; Youssef et al., 2015a), weights of evidence (WOE) (Chen et al., 2016c; Oh and Lee, 2011; Ozdemir and Altural, 2013; Sharma and Kumar, 2008), evidential belief function (EBF) (Pradhan et al., 2014; Tien Bui et al., 2013), certainty factor (CF) (Chen et al., 2016d; Devkota et al., 2013; Kanungo et al., 2011), analytical hierarchy process (AHP) (Chen et al., 2016d; Demir et al., 2013; Shahabi et al., 2014; Yalcin et al., 2011), logistic regression models (Costanzo et al., 2014; Devkota et al., 2013; Lee et al., 2007; Nourani et al., 2014; Ozdemir and Altural, 2013), and multiple logistic regression models (Felicísimo et al., 2013; Lee, 2007; Ohlmacher and Davis, 2003).

Because the prediction capability of these proposed models are critical, machine learning models have also been investigated, such as fuzzy

* Corresponding author.
E-mail addresses: chenwei.0930@163.com (W. Chen), honghaoyuan333@gmail.com, hong_haoyuan@outlook.com (H. Hong).

http://dx.doi.org/10.1016/j.catena.2016.11.032
0341-8162/

logic (Guettouche, 2013; Pourghasemi et al., 2012; Pradhan, 2010; Sharma et al., 2013), fuzzy rule-based classifier (Pham et al., 2016; Tien Bui et al., 2014), neuro fuzzy (Dehnavi et al., 2015; Pradhan, 2013), multivariate adaptive regression splines (MARS) (Conoscenti et al., 2015; Felicísimo et al., 2013; Vorpahl et al., 2012; Wang et al., 2015a), neural networks (Lee et al., 2007; Park et al., 2013; Tien Bui et al., 2016c; Yilmaz, 2010), fuzzy k-nearest neighbor (Tien Bui et al., 2016a), Naïve Bayes (Tsangaratos and Ilia, 2016), support vector machines (Chen et al., 2016b; Colkesen et al., 2016; Xu et al., 2012), least-squared support vector machines (Tien Bui et al., 2016b), and relevant support vector machines (Hoang and Tien Bui, 2016).

A literature review shows that each machine learning model has its strengths and weaknesses, and in general, its behavior depends on characteristics of different study areas. Therefore, comparisons of machine learning models for landslide susceptibility assessment are highly desired. Although several comparison works have been carried out by researchers, such as Pradhan (2013), Hong et al. (2015), Youssef et al. (2015b), and Tien Bui et al. (2016c). However, there are still some state-of-the-art models, such as logistic model tree (LMT), random forest (RF), and classification and regression tree (CART), which have been rarely employed for landslide susceptibility assessment, and therefore they should be further investigated and compared. We address these investigations here by applying, verifying, and comparing three machine learning techniques LMT, RF, and CART for landslide susceptibility mapping, with a case study at the Long County area. Twelve landslide conditioning factors were considered using these three models in GIS. The results were validated using the area under the receiver operating characteristic (ROC) curve method and statistical measures.

2. General situation of the region

The study area (Long County) is located in Shaanxi Province, China, within latitudes 34°35′17″ N to 35°6′45″ N, and longitudes 106°26′32″ E to 107°8′11″ E (Fig. 1). The study area land use types are mainly farmland, bare land, residential areas, water, forest, and grass. The altitude ranges from 778 m to 2467 m, and decreases from west to the east. Qian River and Wei River are the main rivers in the study area, both belonging to the Yellow River network.

According to a Shaanxi Province Meteorological Bureau (http://www.sxmb.gov.cn) report, the study area has a warm temperate continental monsoon climate, with average annual temperature of approximately 10.7 °C and annual rainfall of approximately 600 mm. The average annual evaporation is 1363 mm, and average relative humidity of approximately 70%. The average number of days with precipitation is 120. The rainy season is from May to September, accounting for 75.4% of yearly rainfall. Average annual wind speed 1.5 m/s, and maximum wind speed is 8.4 m/s.

The study area is located at the borders of the southern margin of the Ordos syncline and Qin-Qi geosyncline. The strata are mainly Mesoproterozoic, Cambrian, Ordovician, Triassic, Cretaceous, Neogene, and Quaternary. There are three major faults that divide the study area into distinct structural zones, including (1) the Guguan-Badu (NW–SE direction), (2) the Xinjichuan-Yabo (NW-SE direction), and (3) the Taoyuan-Guichuansi (NW-SE direction). The main lithologies in the study area are loess, mudstone, sandstone, conglomerate, glutenite, limestone, and igneous rocks (Fig. 2).

3. Materials and methods

In the current study, a digital elevation model (DEM) with 30 × 30 m resolution was used to extract a set of topographic factors. The DEM was provided by the International Scientific & Technical Data Mirror Site, Computer Network Information Center, Chinese Academy of Sciences, and available at http://www.gscloud.cn. LANDSAT-8 satellite images with 30 × 30 m spatial resolution were also provided by the same institution. Study area lithology maps at a scale of 1:200,000 were collected from the local Land and Resources Bureau. Meteorological data was collected and compiled from the government of Meteorological Bureau of

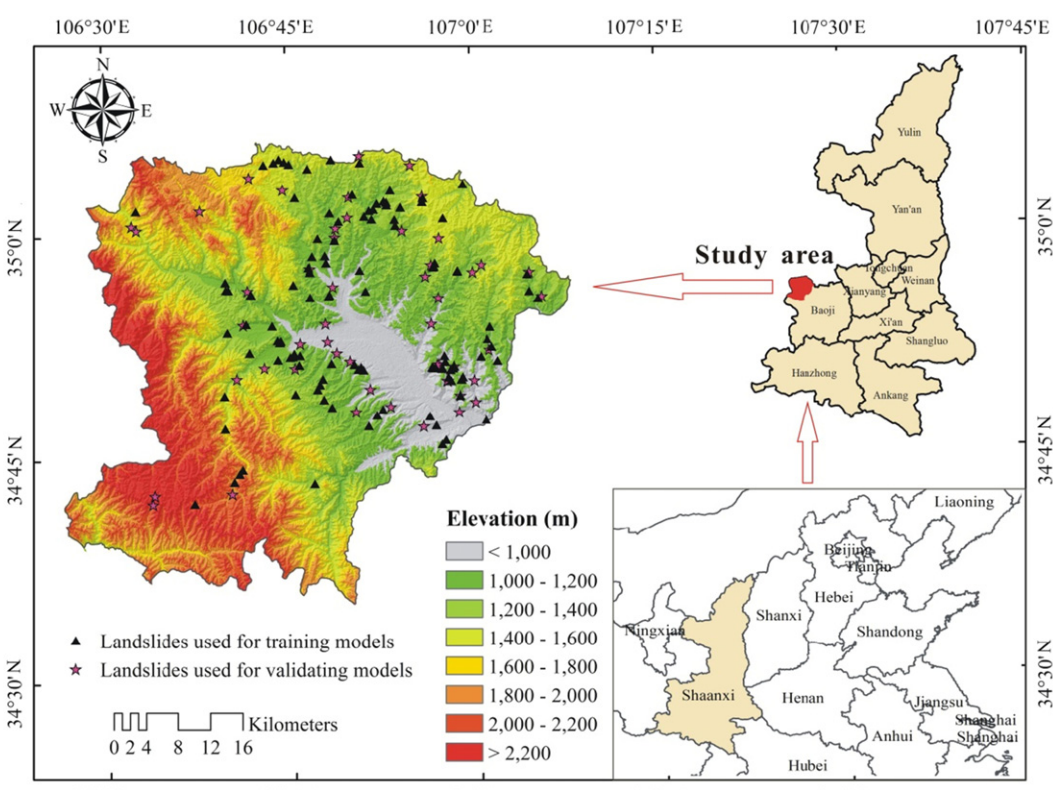

Fig. 1. Study area location of the study area and landslide inventory map.